
Nucleotide sequence and genome organization of carnation mottle virus RNA

H.Guilley, J.C.Carrington⁺, E.Balázs, G.Jonard, K.Richards and T.J.Morris⁺

Laboratoire de Virologie, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue Descartes, 67084 Strasbourg Cédex, France, and ⁺Department of Plant Pathology, University of California, Berkeley, CA 94720, USA

Received 31 July 1985; Accepted 23 August 1985

ABSTRACT

The complete nucleotide sequence of carnation mottle genomic RNA (4003 nucleotides) is presented. The sequence was determined for cloned cDNA copies of viral RNA containing over 99 % of the sequence and was completed by direct sequence analysis of RNA and cDNA transcripts. The sequence contains two long open reading frames which together can account for observed translation products. One translation product would arise by suppression of an amber termination codon and the sequence raises the possibility that a second suppression event could also occur. Sequence homology exists between a portion of the carnation mottle virus sequence and that of putative RNA polymerases from other RNA viruses.

INTRODUCTION

Carnation mottle virus (CarMV) contains a single-component, plus-sense RNA genome of approximately 4.0 kilobases (kb) (1,2). This isometric virus infects a wide range of host plants and is one of the simplest of all eukaryotic infectious agents with regard to genome complexity. The small genome size renders CarMV highly accessible to molecular genetic analysis ; thus, CarMV may serve as a useful model to understand RNA plant virus genome structure, function, and regulation.

The structure and coding properties of CarMV RNA have been subjects of recent study. In addition to genome-length RNA, two major subgenomic species are produced in vivo and encapsidated in virions (3). Both molecules (previously estimated to have lengths of 1.6 and 1.75 kb) are derived from, and apparently colinear with, the 3' end of the viral genome. Contrary to an earlier study using a wheat-germ extract translation system (4) the small subgenomic RNA exhibits potent messenger activity for production of capsid protein (38,000 molecular weight) in reticulocyte lysates (5,6). The role of the 1.75 kb RNA remains unclear, although it may possess weak messenger activity for coat protein production. Like several other plant viruses, the coat protein gene remains silent, or nearly silent, on full-length genomic

RNA. In contrast, the 4.0 kb genomic RNA programs at least two products in reticulocyte lysates, with apparent molecular weights of 30-34,000 and 77-88,000 (5,6). Experiments involving limited proteolysis of these two polypeptides with α -chymotrypsin suggest they share common amino acid sequences (6). The precise nature of this relationship, however, is unresolved in view of the diverse mechanisms of gene expression exhibited by RNA plant viruses (see 7 for review).

Assuming p30-34 and p77-80 originate from the same translational unit, we can account for 80-90 % of the CarMV coding capacity. To more clearly understand CarMV RNA structure and relationships among gene products, as well as the relationship of CarMV among other RNA viruses, the complete nucleotide sequence has been determined. This was facilitated by the availability of complementary DNA clones representing 99.5 % of the genome (Fig. 1). We have identified two long open reading frames which account for the observed translation products. The sequence suggests derivation of at least one product by an amber-termination read-through event. In addition, a region of amino acid sequence homology has been detected between the CarMV read-through domain and putative RNA-dependent polymerases of several other plant and animal viruses (8,9).

MATERIALS AND METHODS

Virus and RNA

Carnation mottle virus (isolate CarMV-B) was propagated on Chenopodium quinoa purified and extracted for RNA as described earlier (3).

cDNA clones

Construction of clone pCarMV-1C (nt 336-4003 ; Fig. 1) has been described previously (3). The clone pCXT-17 (nt 22-332, Fig. 1) was obtained by primer extension. The primer was a single stranded DNA fragment extending from an EcoRI site at the 5' terminus of the pCarMV-1C cDNA insert to the EcoRV site at nt 477 (Fig. 1). pCarMV-1C DNA (50 μ g) was digested with EcoRI, EcoRV and calf intestinal phosphatase (Boehringer) and the restriction fragment mix was 5' end labelled with polynucleotide kinase (Amersham). The restriction fragments were heat denatured and the small single-stranded EcoRV-EcoRI fragments were separated from one another by electrophoresis in a 10 % polyacrylamide gel. After elution from the gel and alcohol precipitation the fragment complementary to the RNA (identified by sequence analysis) was taken up in 25 μ l sterile water for use as a primer. Primer extension was carried out essentially as described by Gubler and Hoffman (10). First strand

synthesis was performed for 1 h at 42° in a final volume of 50 µl of 100 mM Tris, pH 8.3, 140 mM KCl, 10 mM MgCl₂, 400 µM each dNTP, 50 µCi α³²P-dCTP, 2 mM DTT, 10 µg CarMV genomic RNA, 0.3 µg of EcoRV-EcoRI primer and 50 units reverse transcriptase (Life Sciences). After phenol extraction and alcohol precipitation complementary strands were synthesized in 50 µl final volume of 20 mM Tris, pH 7.6, 100 mM KCl, 10 mM MgSO₄, 50 µg/ml bovine serum albumin, 100 µM each dNTP, 10 units DNA polymerase I (BRL), 0.5 units RNase H (BRL) and about 15 ng cDNA:RNA hybrid. The mix was incubated sequentially for 1 h at 14° and 1 h at 22°. Double stranded cDNA was dC-tailed and cloned into Pst-I cut dG-tailed pUC8 as described before (3).

Sequence analysis

The cDNA insert of pCarMV-1C was isolated as two fragments by preparative gel electrophoresis after digestion of the plasmid with PvuII. The two insert DNA bands were visualized by ethidium bromide staining and extracted from the gel by the method of Vogelstein and Gillespie (11). After secondary digestion (with AhaIII, AvaI, BglII, DdeI, EcoRI, HindIII, HinfI, NcoI, RsaI, Sau3A or TaqI) the restriction fragments were 5'-³²P-end labelled with polynucleotide kinase and fractionated by polyacrylamide gel electrophoresis. If the fragment mix was not too complex the fragments were submitted to strand separation prior to electrophoresis. Otherwise, end-labelled double stranded fragments were purified on a first polyacrylamide gel, denatured and the single strands of each fragment separated on a second gel. The short DNA insert of pCXT-17 was sequenced from polylinker restriction sites flanking the insert. About 95 % of the final sequence of pCarMV-1C and pCXT-17 was characterized on both strands. Sequence data were manipulated with University of Wisconsin Genetics Group software (12) run on a VAX 11/750 computer. Protocols for end labelling, strand separation and other procedures are given by Maxam and Gilbert (13), Franck *et al.* (14), and Maniatis *et al.* (15).

Decapping and 5' end labelling CarMV RNA

CarMV RNA (20 µg) was incubated for 30 min at 37° with one unit of tobacco acid pyrophosphatase (BRL) in 50 mM sodium acetate, pH 6, 1 mM EDTA, 10 mM β-mercaptoethanol. After phenol extraction and alcohol precipitation the resuspended RNA was dephosphorylated by digestion for 30 min at 27° with 0.5 U of calf intestinal phosphatase in 25 mM Tris-HCl, pH 8, 0.2 mM EDTA. Full length RNA was then separated from smaller species by sucrose gradient centrifugation. The genome length RNA was 5'-³²P-end labelled with polynucleotide kinase and the peak of radioactivity corresponding to full-length RNA was purified on a second sucrose gradient. The 5' terminal

sequence of the decapped 5'-end labelled RNA was determined by the wandering spot method after partial P1 nuclease digestion (16) or by partial enzymatic degradation (17).

Synthetic deoxyoligonucleotide primers

The deoxyoligonucleotides 5'-TCCCATAGTGAAAGC-3' and 5'-ATCCACTTCGCCTT-3', complementary to residues 61-75 and 358-372, respectively, of the CarMV sequence, were synthesized by the phosphotriester method. The deoxyoligonucleotides were 5'³²P-end labelled with polynucleotide kinase and purified by polyacrylamide gel electrophoresis. About 100 ng of each 5'³²P-end labelled deoxyoligonucleotide was mixed with 20 µg CarMV RNA in 40 µl 7 mM Tris-Cl, pH 8, 50 mM KCl, 5 mM MgCl₂, 5 mM DTT, heated to 65° for 5 min and allowed to cool to room temperature. The solution was adjusted to 50 µM of each dNTP, 40 units of reverse transcriptase were added and the reaction was incubated at 37° for 30 min. End-labelled cDNA runoff transcripts were separated from nonelongated primer by polyacrylamide gel electrophoresis.

RESULTS AND DISCUSSION

Sequence analysis of cDNA clones

CarMV genomic RNA has an estimated length of 3800-4000 nucleotides (3). In an earlier publication (3) we described pCarMV-1C, a pUC8 recombinant plasmid containing a CarMV cDNA insert of about 3800 base pairs. Because it contains most if not all the RNA sequence pCarMV-1C was chosen as starting material for sequence studies. For most experiments the cDNA insert of pCarMV-1C was isolated as two fragments by preparative agarose gel electrophoresis after digestion with PvuII. This enzyme cuts once within the insert (nt 1034 ; all numbering refers to the final sequence) and twice in the pUC8 vector, once near each end of the insert (3). After secondary digestion of each PvuII fragment with numerous restriction enzymes and sequence analysis (see Materials and Methods) a complete nonambiguous sequence for each PvuII fragment was obtained. The overlap between the two PvuII fragments making up the insert was characterized in a separate experiment by isolating the fragment extending from the BamHI site in the pUC8 polylinker to the insert BglII site (nt 1379) and sequencing across the PvuII site from neighboring TaqI sites.

The pCarMV-1C cDNA insert extended from nucleotide 336 to 4003 (Fig. 1). The insert sequence was flanked by an oligo dG tract of 15 residues at the 5' extremity and a poly dA of approximately 45 residues at its 3' end. For cloning purposes a poly rA tail had been added to the 3' end of CarMV RNA

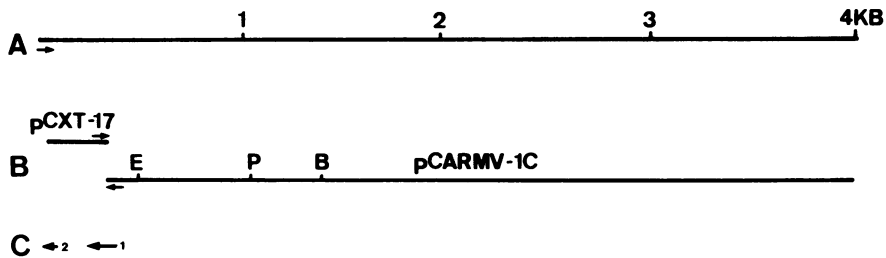


Figure 1. Map of cDNA clones and sequence data for CaMV genomic RNA. (A) Schematic of CaMV genomic RNA showing extent of sequence determined by direct analysis of ^{32}P -labelled RNA. (B) Plasmid cDNA clones of CaMV RNA. The arrow (not to scale) indicates the portion of the sequence inverted in pCARMV-1C (see text). E, P and B denote EcoRV, PvuII and BglII sites in the insert DNA. (C) Sequence analysis of $5' ^{32}\text{P}$ cDNA transcribed from synthetic primers.

with *E. coli* poly A polymerase (3) so that the poly dA tract of the clone serves to define the orientation of the insert sequence with respect to the RNA.

From its size it follows that the pCARMV-1C insert must represent most of the CarMV genome. Comparative sequence analysis of the RNA, however, is necessary to establish whether the cDNA extends all the way to the 5' end. Preliminary experiments showed that genome length RNA could be $5' ^{32}\text{P}$ -labelled to some extent with polynucleotide kinase after incubation with calf intestinal phosphatase. Incorporation was four times greater, however, if the RNA was first treated with tobacco acid phosphatase, indicating that most of the molecules possess a $m^7\text{G}$ cap. Similar results were obtained with RNA which was first decapped chemically and then recapped by treatment with guanylyl transferase in the presence of $\alpha\text{-}^{32}\text{P}$ GTP (KR, personal observations). Genomic CarMV RNA which had been decapped enzymatically and then 5'-end labelled with polynucleotide kinase was purified on a sucrose gradient and sequenced by the wandering spot method. The 5' terminal sequence of 16 nucleotides so obtained (Fig. 2) was extended for an additional ≈ 50 nucleotides with some uncertainties by the partial enzymatic digestion of Donis-Keller *et al.* (17) (data not shown).

The pCARMV-1C insert does not overlap the sequence obtained for the 5' terminus of CarMV genomic RNA by direct sequence analysis. Therefore, cDNA clones extending further toward the 5' extremity were prepared by primer extension. As noted earlier (3) the pCARMV-1C insert sequence begins with an EcoRI site, with the last G of the oligo dG tract donating the first

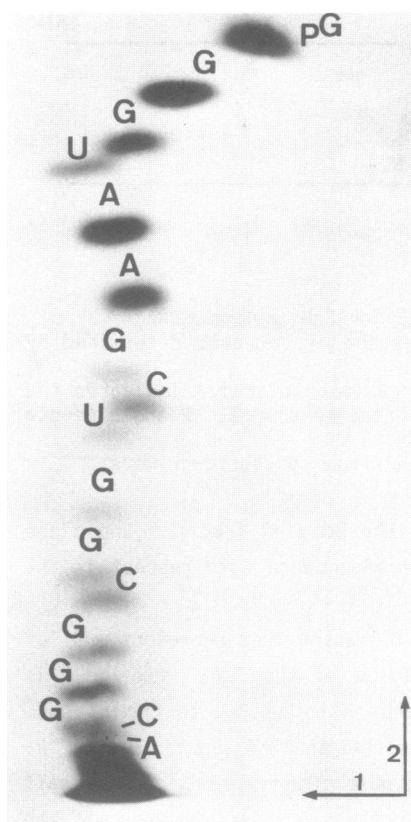


Figure 2. Sequence analysis of decapped 5' 32 P-labelled CarMV genomic RNA by the mobility shift (wandering spot) method after partial P1 nuclease digestion. First dimension : electrophoresis on cellulose acetate at pH 3.5 ; second dimension : homochromatography.

nucleotide of the site (G_{14} GAATTC...). A short minus sense single stranded DNA fragment extending from this EcoRI site to a nearby EcoRV site (nt 477) was prepared from pCarMV-1C, annealed to CarMV RNA and extended with reverse transcriptase. After dC tailing of the cDNA and second strand synthesis the double stranded cDNA was inserted into dG tailed pUC8 and clones were screened for those containing viral sequences. The yield of recombinant clones was considerably below expectation for the amount of starting material used in cDNA synthesis. The recombinant cDNA clone pCXT-17, the clone possessing the longest cDNA insert, starts at the position corresponding to nt 22 of the final sequence, thus overlapping the portion of the sequence established directly with 5' labelled RNA. The 3' terminus of pCXT-17, however, extends only to nt 332 of the final sequence and hence does not overlap the pCarMV-1C sequence (Fig. 1).

cDNA from synthetic primers

In order to characterize the missing sequence between pCXT-17 and pCarMV-1C and to confirm that portion of the 5' terminal sequence determined by direct RNA sequencing methods two synthetic oligonucleotides complementary to residues 358-372 (Primer 1) and 61-75 (Primer 2) were prepared. The synthetic oligonucleotides were 5'-³²P-end labelled, annealed to CarMV RNA and extended with reverse transcriptase. Polyacrylamide gel electrophoresis of the reverse transcripts showed that in each case most of the elongated cDNA molecules extended to the 5' end of the RNA. Full length cDNA was eluted from the gel and sequenced by the partial chemical degradation technique (13).

Sequence analysis of the cDNA transcribed from Primer 1 spanned the gap (of 3 nucleotides) separating pCXT-17 and pCarMV-1C. Furthermore the cDNA sequence revealed that the first 12 residues of pCarMV-1C were not present in the RNA. This block of 12 "extra" residues in pCarMV-1C proved to be an inverted version of the RNA sequence immediately upstream (nt 327-338). Such sequence inversions often occur at the 5' end of cDNA inserts and several models have been put forward suggesting how this type of artefact could arise during cDNA synthesis (18,19). The single-stranded EcoRI-EcoRV DNA fragment used as primer in the construction of pCXT-17 had its 3' terminus in the inverted repeat region of pCarMV-1C. Consequently the primer sequence in this region was not complementary to viral RNA, no doubt accounting for the observed low yield of recombinant clones mentioned above. We suspect that the functional primer in the experiment which produced pCXT-17 was in fact a degradation product of the EcoRI-EcoRV DNA fragment lacking the noncomplementary portion of the sequence.

The sequence of CarMV genomic RNA (4003 nucleotides) deduced from the above results is presented in Fig. 3. Parts of the sequence near the 5' terminus (nt 22-60, 237-332, 336-357) were characterized with both cloned cDNA and by direct sequence analysis of primer elongated cDNA. When they could be compared, the clone and cDNA sequences differed in 14 of 157 positions, as indicated in Fig. 3. 5 of the 11 changes that occur in the coding region (see below) result in amino acid substitutions, several of which are nonconservative. We do not yet know if this degree of sequence polymorphism is characteristic of the entire molecule or is limited to the 5' terminal region. It is worth noting, however, that the virus used for construction of pCarMV-1C and pCXT-17 was prepared in Berkeley while that used for the primer extension experiments was grown up (from the same inoculum) in Strasbourg. Differences in plant growth conditions in the two laboratories may account for at least some of the observed sequence

1 GGGTAAGCTGGCGGGCAACCTACACTCTATCTTACATACATTATATATCTTAAACCATCTGCTTTCACATATGGGATTACCCAGTCTACTAGTGGAAAGGCG
 AT C M G L P S L L V E G V
 101 I G C T L V G C L V A V G S A A L A V R A T I G V V E F N R E C V
 TTATTTGGATCCACTTAGTTGGTGGTTTGGTGGCTGTGGTAGCGCAGCGCTGGCGCTTCCGGGACCATCTCGGTAGTGGAAATCAATAGGAGGTGTGT
 R G A R R I V S S G G R C L V V Q S P V G N P N Q G L I R G E D E
 201 GCGCGGGGCGAGAAGGATGATCTAGCGGAGGGAGGTGCTTAGTAGTCCAATCCCTTATGGCAATCTTAACCAAGGATTGATAAGGGGTGAAGACGAA
 A A A
 E I D N V E E S T P V E L T P L I E V K A E V D G K E V V V S K K R
 301 GAAATTGATTAACGTGGAGGAGTCAACTCCGTAGAGTTCAGTCCGTGATCGAGGTTAAGGCGGAAGTGGATGGGAAAGAGGTGGTTGTCTCTAAGAAGC
 G C G A A A G
 V V N R H L R Q R F V R S I A I E A K N H F G G D I S P S K A N Y
 401 GCGTGGTAACAGGCACCTTGAGCAACAGGATTCGTTCGCTCCATAGCCATCGAAGCCAAAAACCACTTTGGCGGTGATATCAGCCCAAGCAAGGCCAATTA
 L S V S K F L T G K C K E R H V V P A H T R D C V S A A M V L V F
 501 CCTTCTGTCTTCCAAGTTTCTAACCGGGAAATGCAAGGAGAGGCGATGTTGTGCCGCTCACACAAGAGAGCTGTGTGAGCGCTGCCATGGTGTGTGTGTC
 T P D V H E I R M H A G L A S D A A Y G I K I A M H S A S I L N R K G W
 601 ACCCTCAGCTCCAGCAGATCAGAAATGTCAGGAGTGGCGTCTGACCCAGCGTACCGAATAAAGATCCGAGTGGCAAGCATAGTGAACAGGAAAGGT
 C M R L M V N P L D R A R W M E M W C V U N G F D S N K P V T F P
 701 GGTGTGGAGGCTAATGCTAAATCCATTAGATAGGCAAGATGCTGGGAAATGTGGTGGTGTGTAACGGGTTTGACTCCACAAACCGCTCACCTTTCC
 K A G G L F Y L N G V E T K I R R G G H P S V I E V D G Q C P L K
 801 CAAATAGGCGGCTGTCTTACCTCAATGTGTGAGAGCAAAATTCGTCTGGAGGCCACCATCAGTGATTGAGGTAGACGGGCAATGCTCTAGTAA
 E R K L Y V Q N A I T T G Y E Y R V H N H S Y A N L R R G L L E R V
 901 GAGAGAAAACCTACGTACAGATGCCATAACCACTGGTTATGAATACAGGGTCCACAACCAATTCGTACGCAAACTCAGACGAGGGCTCCTTGAGAGAG
 F Y V E R N K E L V S C P Q P E P G S F T K E M G Y L R R R F H R V
 1001 TTTTCTATGTTGAGCGCAACAGGAGTATGTCAGCTGTCCGCAACCTGAACCCGGCAGCTTTAAAGAGATGGGATACCTGCGAGCGAGGTCCATAGAGT
 C G M H T R I S A N D L V D C Y Q G R K R T I V E N A A A S L L D
 1101 GTGTGGCAATCATACCCGAGTCTCGCAATGATTTGGTAGATTTGTTATCAGGGCAGGAACGCCAATTTATGAGAAATCGACAGCGTCCCTACTTGAT
 R A I E R K D G D L K T F I K A E K F N V N L K S D P A P R V I Q P
 1201 AGAGCTATCGAAAGGAAGGATGGGATCTCAAGACCTTTATTAAGCAGAGAAATTCACGTAAATCTGAAAAGTGATCTGCTCTCGGTTATACAGC
 R S P R Y N V E L G R Y L K K Y E H H A Y K A L D K I N G G P T V
 1301 CTAGAGCCCTCGCTACAATGTGGAGTTGGCGGCTACTTAAGAAGTATGAACACCTTACAAAGATCTAGACAGTCTGGGGAGGACCAACAGT
 M K G Y T T E E V A Q H I M S A M N Q F O T P V A I G F D M S R F
 1401 CATGAAGAGATACACTACAGAGGAGGTAGCACAGCACATTTGGAGGCGATGGAATCAATTCAGACACCTGTAGCCATAGGATTGACATGTCAAGATT
 D Q H V S V A A L E F E H S C Y L A C F E G D A H L A N L L K M Q L
 1501 GATCAGCATGTGTCTAGCCGCGCTCGAGTTCGAACATTCATCTGTTGGCCTGTTTGAAGGGGAGCGCTCATCTGCGCAACTTGCTTAAAGATGCAAC
 V M H G V G F A S N G M L R Y T K E G C R M S G G D M N T A L G N C
 1601 TGGTGAATCTGGCGTGGTTTTCGAGCAATGGAATGTGGCGATACACAAGGAAGGTGTCAGATGAGCGGTGACATGAACACTGCCCTGGGCAACTG
 L L A C L I T K L M K I R S L I N N G D D C V L I C E R T D I
 1701 CTGTGTAGCTTGCTTTATCACCAACACTTAATGAAATCAGGAGCAGACTGATCAACAATGGGATGACTGTCTGCTTATTTCGGAAGAACAGACATC
 D Y V V S N L T T G M S R F G F N C I A E E P V Y E M E K I R F C Q
 1801 GACTACGTCTGTGAGTAATTTAAGCAGCGGATGGAATGCAATTCGAATGATGACAGAGAGGCGAGTGTACGAAATGGAAAGATCAGATTCTGCC
 M A P V F D G A G M L M V R D P L V S M S K D S H S L V H W N N E
 1901 AGATGGCACCGGTGTTGATGAGCAGGCTGGCTGATGTCAGGACCCCTTGGTGAGCATGAACAAGGATTCTCACTCCTTAGTGCAITGGAAACACGA
 T N A K Q N L K S V G M C G L R I A G G V P V V Q E F Y Q K Y V E
 2001 AACGAATGCAAAACATGGCTGAAGTCAGTAGGAATGTGCGGTCTGCGGATGCGCGGGGGGTTCCAGTTGTGCAAGAGTTCTACCAAAAAATACGTTGAA
 T A G N V R E N K N I T E K S S S G F F M H A D R A K R G Y S A V S
 2101 ACAGCCGGAATGTGAGAGAGACAAAAATATCACAGAGAAGATGATTCTGGGTCTTTATGATGCGCGATCGCGCTAAGCGGGCTATTCTGCTGTGT
 E V C R F S F V Q A F G I T P D Q Q I A L E G E I R S L T I N T W
 2201 CTGAGGTGTGCCATTAGCTTCTACAGGCAATGGCATCAGCCAGACAGCAGATCGCTTGGAGGTTAGATCAGGTCTCTCACTATCAACACCA
 V G P Q C E A A D S L W I L N R K Y Q * L E S K C S L G I E E N K
 2301 CGTGGGGCCCCAGTGTGAAGCGGCAGATTCACTATGGATATTGAATCGAAGTACCAGTAGTTGGAAAGCAAAATGCTCGTGGGAATAGAGGAAAAACAA
 R H V D R W P R M P S A N L H L I V L T G V I G L M L L I R L R C T
 2401 AGACACGTGATGCGTGGCAGGATGCCATCCGCAACCTGCACTGATAGTACTAACGGGGGTAATTTGGTTAAATGTTGTCTGATAGAGATTGAGGTGCA
 F T S T F S L P L V L T L N Q I A L S F T G C L L L N S I S R A E
 2501 CATCTACTTCAACTTTAGTTTGGCCCCCATGTAATCTTAATCAGATATAGCGTGTGCTATTTGTGGTCTTCTATTAAGACAGCATCTCTCGGCGAGA
 M E N K G E K I A M W
 2601 R A C Y Y N Y S V D S S K Q Q H I S I S T P N G K A
 AAGACCTTCTATTACAACTACTCTGTGTAGATTCTAAACAACAACACATTTCAATAGTACACCAATGGAATAAAGGAGAAAGATCCCGATGAA
 P T V Q T L A Q K G D K L A V K L V T R G W A S L S T N Q K R R A
 2701 TCCACTCTGTGCAAACTATGGCGAGAGGGGGACAAGTATGCGCTGAAGTGTGTGACAAGAGGTTGGGCTCTCTGAGTACCAAGACAGAGAGAGAGCT
 E M L A G Y T P A I L A F T P R R P R M T N P P P R T S R N S P G Q
 2801 GAAATCTTCTGTGATACATCCAGCGATCTTAGCCTTACACCCCGACGACCGGATGACGAACCTCTCTCAAGAACAGTAGGAATTCACGAGAC

2901 A G K S M T N S K T E L L S T V K G T T G V I P S F E D N W V V S P
AAGCTCGAAAGTCCATGACGATGATGAAGACCGAACTATTAAAGCCGTCAAAGGTACCAACGGGTGTTATCCCAAGCTTTGAAGACTGGGTCTGTTCCCC

3001 R N V A V F P Q L S L L A T T A C C N K Y R I T A L T T K V S P A C S
CCGAAACGTAGCCGCTCTCCCTCAACTCTCGCTGCTGGCGACGAACTCAACAAGTACCGCATTAATCGCGCTACTGTGAAGTACTCACCCGGCGTGCAGC

3101 F E T N G R V A L A L G F N D D A S T T P P T T K V G F Y V D L G K H V E
TTGGAACCAATGGGAGGCTGGCTCTGGGATTCAACGATGACGCTTCGGACACCCCAACCAACGAAGTTGGATTTTACGATTCGGGCAAACTAGTGG

3201 T A A Q T A C K D L V I P V D G K T R F I R D S A S D D A K L V D F
AAACTCTGCACAGACGCTAAGGATCTAGTGATACCAAGTAGACGGCAAAACCCGGTTCTACAGGGACTCGGCTAGTGATGATGCCAAACTAGTCGATT

3301 G R I V L S T Y G F D K A D T V V G E L F I Q Y T I V L S D P T K
TGGACGAATAGCTCTGTCAACATACCGGTTTGCAAGGCTGACACTGTGCTGGCGGAATTGTTTATACAGTACACGATTGTGCTGAGTGACCAACCAAG

3401 T A K I S A G S A N D K V S D G P T Y V V P S V N G N E L Q L R V V A
ACGGCCAAAATTTCACAGCAAGCAACGATAGGTGTCCGACGGCAACGATATGGTCCCTTCGGTTAATGGCAACGAGCTACAACTAAGAGTGGTAG

3501 A G K W C I I V R G T V E G G F T K P T L I G C P G I S G D V D Y E
CCGCTGGGAAATGGTGATTATAGTGCAGGGGTACGGTTGAGGAGGGCTTCAACAAACCCACACTTATGGCGCGGAATACGGCGTGATGTGGAGTATGA

3601 S A R P I A V C E L V T Q M E G Q I L K I T K T S A E Q P L Q W V
AAGTGCACGACCTATCCCGGTTGTGAATTGGTGACAAATAGGAGGCCAGATATGAAAATACCAAGACCTCAGCAGACAACCACTCCAATGGGTT

3701 V Y R M *
GTTTATAGGATGAATGCCAANTGAAGAAAGTTTGTAGACACACCGCGGGAGCTGTGCGCGAGCTGCCACTCCGCCCAAGATGGATGATATTAAACCA

3801 T C T T T G G C A A C G G T G C G T G G C T A A T C G A A C T A G T A A T T G C A T A G C A T C C G A C T G T A A C C T G T T C T C T G C T A A T A C G A G G G T G G C C T G T G A T C C C A

3901 A A C C A T A A A A G A A A C C A A C T T G T T G T G G A A G G C C T G G A A G T A G G A T C A A C C A C T G T T T A T T A A T T A G G G G T A G A G A T T A C T A C T G T A C T C T T T C C C G

4001 CCC 4003

Figure 3. Nucleotide sequence of CarMV genomic RNA written as DNA. Termination codons for the major open reading frames are denoted by stars. Amino acids encoded by the major open reading frames are given in one-letter code over the first base of each codon in the frame. Variations from the cloned cDNA sequence observed when analyzing cDNA transcripts are indicated beneath the sequence.

differences by selection of subvariants from the population.

Coding capacity of CarMV genomic RNA

As noted in the Introduction there is evidence for three major CarMV cell-free translation products : p38 (the viral coat protein), p30-34 and p77-80. Synthesis of p38 is directed by one or perhaps two encapsidated subgenomic RNAs derived from the 3' end of the viral genome while synthesis of p30-34 and p77-80 is directed by genomic RNA. p30-34 and p77-80 have overlapping sequences but the time course of their appearance during protein synthesis rules out the possibility that the smaller protein derives from the larger by proteolytic processing (6). An alternative possibility is that p77-80 arises by partial readthrough of the p27-30 stop codon, a strategy for gene expression which is used by a number of other plant viruses (20).

The CarMV sequence contains two long open reading frames (Fig. 4). The first coding region begins at AUG(70), the first potential initiation codon is the RNA, and is interrupted by a UAG triplet at nt 805 to give a polypeptide with predicted M_r of 26841. An in-phase coding sequence extends beyond this amber termination codon for an additional 1551 nucleotides before reaching a second UAG at nt 2359. A readthrough protein beginning at AUG (70) and terminating at UAG (2359) has M_r of 85831 and we suggest that the 26841

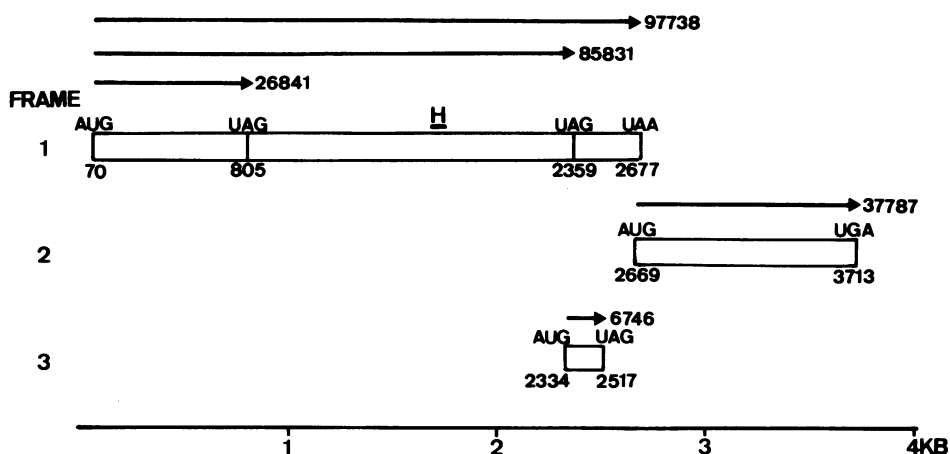


Figure 4. Major open reading frames in CarMV genomic RNA. Hollow boxes represent coding regions with initiation codons and termination codons at the indicated positions. Arrows above the reading frames represent possible translation products. Numbers at the end of each arrow refer to the M_r of each polypeptide as calculated from its amino acid composition. The M_r 85831 and 97738 polypeptides would arise from readthrough of amber termination codons (see text). The black bar labelled H represents the portion of the CarMV polypeptide sequence displaying homology with putative RNA polymerases from other RNA viruses (see Fig. 5).

and 85831 polypeptides correspond to the p30-34 and p77-80 *in vitro* translation products. Readthrough would involve suppression of an amber termination codon as occurs during translation of tobacco mosaic virus genomic RNA (21,22) and beet necrotic yellow vein virus (23, and further observations). Natural suppressor tRNA activity which incorporates tyrosine in place of an amber termination codon has been detected in tobacco (22). Following the UAG triplet at nt 2359 the reading frame extends for an additional 315 nucleotides to a UAA triplet at nt 2677. A double readthrough protein extending from AUG (70) to this ochre termination codon would have M_r of 97738. No such polypeptide was detected in our previous *in vitro* translation experiments but we would not want to rule out the possibility of its synthesis under conditions optimized to promote readthrough.

The second long coding region in the CarMV sequence begins with AUG (2669) and ends with UGA (3713) (Fig. 4). The size of the predicted polypeptide (M_r 37787) and its amino acid composition are similar to viral coat protein (24). Its 3'-proximal position on the genome is consistent with its identity as capsid protein as suggested by cell-free synthesis data (6).

completely lacks a P domain (27).

No other open reading frame exists in the CarMV sequence which could potentially encode a product of greater than 7000 molecular weight. We are faced, however, with the presence of the 1.75 kb, 3' proximal subgenomic RNA found in infected tissue and which presumably functions as a messenger. If we assume co-linearity of genomic and subgenomic species starting at their 3' ends (3), the 1.75 kb RNA 5' terminus should extend into the 85831 coding region. The first open reading frame on this subgenomic RNA should initiate at AUG (2334), which appears in optimal sequence context according to Kozak (28), and terminate at UAG (2517). The predicted translation product consists of 61 amino acid residues (6746 M_r). We have observed small quantities of an in vitro translation product with estimated 10,000 molecular weight primarily stimulated by the 1.75 kb RNA using uniformly-labeled ^3H -amino acids (unpublished), but we are not certain it is derived from the above mentioned open reading frame.

The minus strand of CarMV genomic RNA has limited coding capacity. The longest coding region is 113 amino acids long (nt 3519-3178) and is located in the part of the RNA molecule corresponding to the coat protein cistron. Open reading frames of comparable length also occur in the coat protein region of the minus strands of alfalfa mosaic virus and brome mosaic virus RNA (29) but the significance, if any, of these potential coding regions is unknown.

Sequence homology with other viruses

As more sequence information about plant viruses becomes available it has been possible to detect sequence homology between coding regions of viruses which are not classified together by conventional criteria (e.g. 8,30,31). Such comparisons not only provide insight into virus evolution but can also furnish clues as to protein function. Kamer and Argos (9) have recently aligned polypeptide sequences corresponding to the putative viral coded RNA-dependent RNA polymerase from eight different plant and animal RNA viruses. The most highly conserved regions are a gly-asp-asp (GDD) triplet surrounded by hydrophobic amino acids and a second 11 amino acid stretch with extensive homology located 22-37 amino acids upstream of the GDD sequence. A comparable sequence motif occurs in the putative RNA polymerase of the insect virus black beetle virus (32). Figure 5 shows that these conserved sequences are also present in the readthrough portion of the CarMV 85831 polypeptide suggesting that this portion of the CarMV genome encodes a similar function, presumably an RNA polymerase activity.

The primary structures of several (+) sense RNA virus genomes are now known. Of those sequenced, CarMV shares gene organizational features most noticeably with TMV. Each possesses 1) a single-component genome, 2) a readthrough protein initiating at the 5'-proximal open reading frame, and 3) at least two 3'-derived subgenomic RNAs of which the smallest encodes coat protein. They also have 5' leader sequences of 68 (TMV) and 69 (CarMV) bases substantially devoid of G residues (33). Additionally, the polypeptide context in which homology occurs appears similar in both (i.e., read-through domains). It is therefore interesting how CarMV and TMV dramatically diverge when compared with BMV, AMV, and Sindbis. Tobacco mosaic virus and these latter viruses share three domains of amino acid sequence homology among non-structural protein (34 ; 8). Carnation mottle virus shares one obvious homology as mentioned above. There is also same homology between the CarMV pre-readthrough (26841) polypeptide and the C-terminally located one of the two other conserved regions in TMVp126, BMV and AMV RNA 1-encoded proteins, and Sindbis nsP1 and nsP2 (data not shown). Given the size of these homologs (\approx 500 residues in each), however, CarMV could, at most, only contain sequences similar to one and one-half of the three domains.

Interestingly, TMV, AMV, BMV, and other plant RNA viruses harbor genes for another non-structural polypeptide of around 30,000 M_r . Defective TMV 30 K protein correlates with deficient cell-to-cell movement capability (35). Each of these genes reside at a locus immediately upstream from their respective coat protein coding sequences. Carnation mottle virus does not code for a gene product of this size at a locus comparable to the other viruses. It is possible that the postulated 6746 M_r polypeptide (see above), encoded by sequences just upstream from the coat protein gene and potentially expressed from the 1.75 Kb subgenomic RNA, serves a role similar to that of one or more of the 30 K proteins.

It is safe to say a relatively diverse group of viruses (e.g., TMV, tripartite viruses, animal alphaviruses) have evolved a set of structurally and functionally similar proteins, but what about CarMV? With less than one-half of the non-structural protein coding capacity of TMV, for example, it might simply require and perform fewer biochemical steps to replicate and spread in infected plants. Alternatively, CarMV replication may be as or more complex relative to the others, but it has adapted host proteins as functional substitutes for viral-encoded polypeptides. The economy of CarMV non-structural proteins might also be explained by assuming multifunctionality, or perhaps CarMV encodes as many functional domains as the oth12ers but on a

"bare minimum" of protein. At this point, we do not know which if any of these scenarios is (are) valid, certainly a reflection of our limited knowledge of these viruses at the genetic and biochemical levels.

REFERENCES

1. Waterworth, H.E. and Kaper, J.M. (1972) Purification and properties of carnation mottle virus and its ribonucleic acid. *Phytopathology* 62, 959-964.
2. Kaper, J.M. and Waterworth, H.E. (1973) Comparison of molecular weights of single-stranded viral RNAs by two empirical methods. *Virology* 51, 183-190.
3. Carrington, J.C. and Morris, T.J. (1984) Complementary DNA cloning and analysis of carnation mottle virus RNA. *Virology* 139, 22-31.
4. Salomon, R., Bar-Joseph, M., Soreq, H., Gozes, I. and Littauer, U.Z. (1978) Translation in vitro of carnation mottle virus RNA. Regulatory function of the 3'-region. *Virology* 90, 288-298.
5. Harbison, S.A., Wilson, T.M.A. and Davies, J.W. (1984) An encapsidated, subgenomic messenger RNA encodes the coat protein of carnation mottle virus. *Bioscience Reports* 4, 949-956.
6. Carrington, J.C. and Morris, T.J. (1985) Characterization of the cell-free translation products of carnation mottle virus genomic and subgenomic RNAs.
7. Davies, J.W. and Hull, R. (1982) Genome expression of plant positive-strand RNA viruses. *J. Gen. Virol.* 61, 1-14.
8. Haseloff, J., Goelet, P., Zimmer, D., Ahlquist, P., Dasgupta, R. and Kaesberg, P. (1984) Striking similarities in amino acid sequence among non-structural proteins encoded by RNA viruses that have dissimilar genomic organizations. *Proc. Natl. Acad. Sci. USA* 81, 4358-4362.
9. Kamer, C. and Argos, P. (1984) Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. *Nucleic Acids Res.* 12, 7269-7282.
10. Gubler, V. and Hoffman, B.J. (1983) A simple and efficient method for generating cDNA libraries. *Gene* 25, 263-269.
11. Vogelstein, B. and Gillespie, D. (1979) Preparative and analytical purification of DNA from agarose. *Proc. Natl. Acad. Sci. USA* 76, 615-619.
12. Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12, 387-395.
13. Maxam, A.M. and Gilbert, W. (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* 65, 499-560.
14. Franck, A., Guillely, H., Jonard, G., Richards, K. and Hirth, L. (1980) Nucleotide sequence of cauliflower mosaic virus DNA. *Cell* 21, 285-294.
15. Maniatis, T., Fritsch, E.F. and Sambrook, J. *Molecular Cloning : A Laboratory Manual*. Cold Spring Harbor Laboratory (1982).
16. Richards, K.E., Jonard, G., Guillely, H. and Keith, G. (1977) Leader sequence of 71 nucleotides devoid of G in tobacco mosaic virus RNA. *Nature* 267, 548-550.
17. Donis-Keller, H., Maxam, A.M. and Gilbert, W. (1977) Mapping adenines, guanines and pyrimidines in RNA. *Nucleic Acids Res.* 4, 2527-2538.
18. Fields, S. and Winter, G. (1981) Nucleotide sequence heterogeneity and sequence rearrangements in influenza virus cDNA. *Gene* 15, 207-214.
19. Volckaert, G., Tavernier, J., Derynck, R., Devos, R. and Fiers, W. (1981) Molecular mechanisms of nucleotide-sequence rearrangements in cDNA clones of human fibroblast mRNA. *Gene* 15, 215-223.
20. Joshi, S. and Haenni, A.L. (1984) Plant RNA viruses : strategies of expression and regulation of viral genes. *FEBS Lett.* 177, 163-174.

21. Pelham, H.R.B. (1978) Leaky AUG termination codons in tobacco mosaic virus RNA. *Nature* 272, 469-471.
22. Beier, H., Barciszewska, M., Krupp, G., Mitnacht, R. and Gross, H.J. (1984) UAG readthrough during TMV RNA translation : isolation and sequence of two tRNAs^{Tyr} with suppressor activity from tobacco plants. *EMBO J.* 3, 351-356.
23. Ziegler, V., Richards, K., Guilley, H., Jonard, G. and Putz, C. (1985) Cell-free translation of beet necrotic yellow vein virus : readthrough of the coat protein cistron. *J. Gen. Virol.* in press.
24. Tremaine, J.H. and Goldsack, D.E. (1968) The structure of regular viruses in relation to their subunit amino acid composition. *Virology* 35, 227-237.
25. Hopper, P., Harrison, S.C., and Sauer, R.T. (1984) Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. *J. Mol. Biol.* 177, 701-713.
26. Harrison, S.C. (1983) Virus structure : high-resolution perspectives. *Adv. Virus Research* 28, 175-240.
27. Abad-Zapatero, C., Abdel-Meguid, S.S., Johnson, J.E., Leslie, A.G.W., Rayment, I., Rossman, M.G., Suck, D., and Tsukihara, T. (1980) Structure of southern bean mosaic virus at 2.8 Å resolution. *Nature (London)* 286, 33-39.
28. Kozak, M. (1984) Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. *Nature* 308, 241-246.
29. Van Vloten-Doting, L., Dubelaar, M. and Bol, J.F. (1982) Open reading frame in the minus strand of two plus type RNA viruses. *Plant Molec. Biol.* 1, 155-158.
30. Franssen, H., Leunissen, J., Goldbach, R., Lomonosoff, G. and Zimmern, D. (1984) Striking similarities in amino acid sequence among non-structural proteins encoded by RNA viruses that have dissimilar genomic organization. *Proc. Natl. Acad. Sci. USA* 81, 4358-4362.
31. Toh, H., Hayashida, H. and Miyata, T. (1983) Homology of reverse transcriptase of retrovirus with putative polymerase gene products of hepatitis B virus and cauliflower mosaic virus. *Nature* 305, 827-829.
32. Dasmahapatra, B., Dasgupta, R., Ghosh, A. and Kaesberg, P. (1985) Structure of the black beetle virus genome and its functional implications. *J. Mol. Biol.* 182, 183-189.
33. Richards, K., Guilley, H., Jonard, G., and Hirth, L. (1978) Nucleotide sequence at the 5' extremity of tobacco-mosaic-virus RNA. *Eur. J. Biochem.* 84, 513-519.
34. Ahlquist, P., Strauss, E.G., Rice, C.M., Strauss, J.H., Haseloff, J., and Zimmern, D. (1985) Sindbis virus proteins nsP1 and nsP2 contain homology to nonstructural proteins from several RNA plant viruses. *J. Virol.* 53, 536-542.
35. Leonard, D.A., and Zaitlin, M. (1982) A temperature-sensitive strain of tobacco mosaic virus defective in cell-to-cell movement generates an altered viral-coded protein. *Virology* 117, 416-424.